Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
13.4.2022
J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Allowed equipment: writing tools, calculator (symbolic and graphic OK), at most A4-size cheat sheet written on one side.

P1 In each question, answer TRUE or FALSE. In this problem you do not need to write reasons. 0.5 points per item for correct answer.

(a) Descriptive statistics aims to draw conclusions about a population based on a sample.

(b) Sample mean is the 0.5-quantile of a sample.

(c) If a distribution is symmetric around zero, it is called the normal distribution.

(d) The chi-squared (χ^2) distribution is symmetric around zero.

(e) Positive skewness indicates that the distribution has a long right tail.

(f) In linear regression, correlation coefficient is the slope of the regression line.

(g) Confidence level indicates how much confidence you have in the model assumptions.

(h) In a simple linear regression $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the error term ϵ_i is usually assumed to have zero expected value.

(i) Bootstrapping is a method for using linear regression with multiple predictor variables.

(j) Signed rank tests make stricter assumptions than sign tests.

(k) In hypothesis testing, the null hypothesis is rejected when the p-value is greater than the significance level.

(l) In multiple testing, Bonferroni correction increases the probability of Type II errors.

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.

H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give their general definitions but explain what they mean here specifically). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name and briefly explain two ways besides Q-Q plot for checking/testing the normality of a sample. **(2p)**

Statistical inference
MS-C1620
Department of Mathematics and Systems Analysis
Aalto University

Exam
13.4.2022
J Kohonen

P3 (a) Draw a scatter plot of two variables that have:

(i) perfect linear dependence **(1p)**

(ii) perfect monotonic dependence but not perfect linear dependence **(1p)**

(b) Is it possible for two variables to have perfect linear dependence but not perfect monotonic dependence? Explain why or why not. **(2p)**

(c) Explain Spearman's rank correlation coefficient. When is it used and how? **(2p)**

P4 Consider a dataset of $n = 7$ observations with x values

$$3.0, 3.5, 4.0, 5.0, 5.5, 6.0, 9.0$$

and y values

$$0.0, 0.0, 0.0, 1.0, 1.0, 1.0, 2.0.$$

Let K be the triangular kernel function

$$K(u) = \begin{cases} 1 - |u| & \text{if } |u| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Calculate the Nadaraya-Watson regression function at points $x = 4.0, 4.1, 4.5, 4.9, 5.0$ and 7.5 using K as the kernel. Give at least three decimals. If at some points it cannot be calculated, explain why. **(2p)**

(b) What happens with this dataset if we use the kernel function $K_{0.01}(u) = K(u/0.01)$? Where exactly can the regression be calculated and what are its values there? **(1p)**

(c) Calculate the regression function at $x = 4.0$ and at $x = 7.5$ using the kernel function $K_{100}(u) = K(u/100)$. Give at least three decimals. **(1p)**

(d) Explain in general terms (not just for this dataset) how bandwidth affects Nadaraya-Watson kernel regression. Consider small, intermediate and large values. **(1p)**

(e) Explain how cross-validation can be used to select bandwidth in kernel regression. **(1p)**